Rethinking Spiking Self-Attention Mechanism: Implementing α-XNOR Similarity Calculation in Spiking Transformers

Supplementary Material

A. Main Theorems and Proofs

A.1. Proof of Theorem 1

Theorem 1. *Let $\boldsymbol{q}, \boldsymbol{k} \in \{0,1\}^d$ be independent spike trains with firing rates $f_{\boldsymbol{q}}$ and $f_{\boldsymbol{k}}$, respectively. The similarity scores for dot-product and XNOR similarity both follow a binomial distribution:*

$$\mathrm{Sim}_{\mathrm{DP}}(\boldsymbol{q}, \boldsymbol{k}) \sim \mathrm{Bin}(d, p_1), \tag{31}$$

$$\mathrm{Sim}_{\mathrm{XNOR}}(\boldsymbol{q}, \boldsymbol{k}) \sim \mathrm{Bin}(d, p_1 + p_2), \tag{32}$$

where $p_1 = f_{\boldsymbol{q}} f_{\boldsymbol{k}}$, and $p_2 = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})$, represent the probability of matching spikes and non-spikes, respectively.

Proof. Let $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$ and $\boldsymbol{k} = (k_1, k_2, \ldots, k_d)$ be independent binary vectors of length d, where each $q_i, k_i \in \{0,1\}$. The firing rates are defined as $f_{\boldsymbol{q}} = P(q_i = 1)$ and $f_{\boldsymbol{k}} = P(k_i = 1)$ for all i.

For each position i, since q_i and k_i are independent Bernoulli random variables, we have:

$$P(q_i = 1) = f_{\boldsymbol{q}}, \quad P(q_i = 0) = 1 - f_{\boldsymbol{q}}, \tag{33}$$

$$P(k_i = 1) = f_{\boldsymbol{k}}, \quad P(k_i = 0) = 1 - f_{\boldsymbol{k}}. \tag{34}$$

Then we define two random variables for each position i:

$$X_i = q_i k_i \in \{0,1\}, \tag{35}$$

$$Y_i = \begin{cases} 1, & \text{if } q_i = k_i, \\ 0, & \text{if } q_i \neq k_i. \end{cases} \tag{36}$$

Both X_i and Y_i take values in $\{0,1\}$ and depend on q_i and k_i. Since q_i and k_i are independent for all i, the random variables X_i and Y_i are independent across different i.

Define the dot-product similarity and the XNOR similarity as sums of these random variables:

$$\mathrm{Sim}_{\mathrm{DP}}(\boldsymbol{q}, \boldsymbol{k}) = \sum_{i=1}^{d} X_i, \quad \mathrm{Sim}_{\mathrm{XNOR}}(\boldsymbol{q}, \boldsymbol{k}) = \sum_{i=1}^{d} Y_i. \tag{37}$$

Since each X_i and Y_i are independent Bernoulli random variables, their sums $\mathrm{Sim}_{\mathrm{DP}}$ and $\mathrm{Sim}_{\mathrm{XNOR}}$ follow binomial distributions. Next, we compute the success probabilities for X_i and Y_i.

For X_i and Y_i, we have:

$$P(X_i = 1) = P(q_i = 1) \cdot P(k_i = 1) = f_{\boldsymbol{q}} f_{\boldsymbol{k}}. \tag{38}$$

$$\begin{aligned} P(Y_i = 1) &= P(q_i = 0, k_i = 0) + P(q_i = 1, k_i = 1) \\ &= f_{\boldsymbol{q}} f_{\boldsymbol{k}} + (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}}). \end{aligned} \tag{39}$$

Thus, $X_i \sim \mathrm{Ber}(f_{\boldsymbol{q}} f_{\boldsymbol{k}})$, and $\mathrm{Sim}_{\mathrm{DP}}(\boldsymbol{q}, \boldsymbol{k}) \sim \mathrm{Bin}(d, f_{\boldsymbol{q}} f_{\boldsymbol{k}})$. Similarly, $Y_i \sim \mathrm{Ber}\left(f_{\boldsymbol{q}} f_{\boldsymbol{k}} + (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})\right)$, and $\mathrm{Sim}_{\mathrm{XNOR}}(\boldsymbol{q}, \boldsymbol{k}) \sim \mathrm{Bin}\left(d, f_{\boldsymbol{q}} f_{\boldsymbol{k}} + (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})\right)$.

This completes the proof. □

A.2. Proof of Theorem 2

In this subsection, we first establish the distribution of the score under α-XNOR similarity in Theorem 2. Then, we derive its expectation and variance in Corollary 1.

Theorem 2. *Let $\boldsymbol{q}, \boldsymbol{k} \in \{0,1\}^d$ be independent spike trains with firing rates $f_{\boldsymbol{q}}$ and $f_{\boldsymbol{k}}$, respectively. The score calculated by α-XNOR similarity is distributed as:*

$$P(s = k + l\alpha) = \frac{d!}{k!\, l!\, (d - k - l)!} \cdot p_1^k\, p_2^l\, p_3^{d-k-l}, \tag{40}$$

for $0 \leq k \leq d$ and $0 \leq l \leq d - k$, where $p_1 = f_{\boldsymbol{q}} f_{\boldsymbol{k}}$, $p_2 = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})$, and $p_3 = 1 - p_1 - p_2$.

Proof. Consider two independent spike trains $\boldsymbol{q}, \boldsymbol{k} \in \{0,1\}^d$, where each element q_i and k_i is a Bernoulli random variable with probabilities $P(q_i = 1) = f_{\boldsymbol{q}}$ and $P(k_i = 1) = f_{\boldsymbol{k}}$, respectively, for $i = 1, 2, \ldots, d$.

For each index i, the variables q_i and k_i are independent Bernoulli random variables. The joint probabilities of their possible outcomes are determined by these firing rates. Specifically, the probability that both q_i and k_i are equal to 1 is:

$$P(q_i = 1,\, k_i = 1) = f_{\boldsymbol{q}} \cdot f_{\boldsymbol{k}} = p_1, \tag{41}$$

in which case $\alpha(q_i, k_i) = 1$. Conversely, the probability that both q_i and k_i are equal to 0 is

$$P(q_i = 0,\, k_i = 0) = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}}) = p_2, \tag{42}$$

resulting in $\alpha(q_i, k_i) = \alpha$. In all other scenarios where q_i and k_i differ, the probability is

$$\begin{aligned} P(q_i \neq k_i) &= P(q_i = 1,\, k_i = 0) + P(q_i = 0,\, k_i = 1) \\ &= f_{\boldsymbol{q}}(1 - f_{\boldsymbol{k}}) + (1 - f_{\boldsymbol{q}}) f_{\boldsymbol{k}} = p_3, \end{aligned} \tag{43}$$

and $\alpha(q_i, k_i) = 0$ in these cases.

Since the outcomes at each position i are independent, the joint distribution of K, L, and M follows a multinomial distribution:

$$P(K = k, L = l, M = m) = \frac{d!}{k!\, l!\, m!} \cdot p_1^k\, p_2^l\, p_3^m, \tag{44}$$

where $m = d - k - l$, and $0 \leq k, l, m \leq d$.

The total α-Similarity score s can be expressed in terms of K and L as:

$$s = K \cdot 1 + L \cdot \alpha + M \cdot 0 = K + L\alpha. \tag{45}$$

Therefore, the probability that the α-Similarity score equals $s = k + l\alpha$ is:

$$\begin{aligned} P(s = k + l\alpha) &= P(K = k, L = l) \\ &= \frac{d!}{k!\, l!\, (d-k-l)!} \cdot p_1^k\, p_2^l\, p_3^{d-k-l}, \end{aligned} \tag{46}$$

for all valid combinations of k and l such that $0 \leq k \leq d$ and $0 \leq l \leq d - k$.

This completes the proof. □

Building on this distribution, we calculate the expectation and variance of the score under α-XNOR similarity, as presented in the following corollary.

Corollary 1. *Let $\boldsymbol{q}, \boldsymbol{k} \in \{0,1\}^d$ represent independent spike trains with firing rates $f_{\boldsymbol{q}}$ and $f_{\boldsymbol{k}}$, respectively. The expectation and variance of the score, calculated by α-XNOR similarity, are given as follows:*

$$\mathbb{E}[s] = d \cdot (p_1 + \alpha p_2), \tag{47}$$

$$\mathbb{V}[s] = d \cdot \left(p_1 + \alpha^2 p_2 - (p_1 + \alpha p_2)^2\right), \tag{48}$$

where s denotes the score, $p_1 = f_{\boldsymbol{q}} f_{\boldsymbol{k}}$, $p_2 = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})$, and $p_3 = 1 - p_1 - p_2$.

Proof. To calculate the expectation, we have:

$$\mathbb{E}[\text{score}] = \mathbb{E}[K + L\alpha] = \mathbb{E}[K] + \alpha\mathbb{E}[L]. \tag{49}$$

Given the properties of the multinomial distribution:

$$\mathbb{E}[K] = dp_1, \quad \mathbb{E}[L] = dp_2. \tag{50}$$

Thus,

$$\mathbb{E}[\text{score}] = dp_1 + \alpha dp_2 = d(p_1 + \alpha p_2). \tag{51}$$

For the variance, we have:

$$\mathbb{V}[\text{score}] = \mathbb{V}[K + L\alpha] = \mathbb{V}[K] + \alpha^2\mathbb{V}[L] + 2\alpha\mathbb{C}\text{ov}(K, L). \tag{52}$$

Since K and L are counts of disjoint events in the multinomial distribution, they are negatively correlated:

$$\text{Cov}(K, L) = -dp_1 p_2. \tag{53}$$

Therefore,

$$\mathbb{V}[\text{score}] = dp_1(1 - p_1) + \alpha^2 dp_2(1 - p_2) + 2\alpha(-dp_1 p_2). \tag{54}$$

Simplifying, we obtain:

$$\mathbb{V}[\text{score}] = d\left(p_1 + \alpha^2 p_2 - (p_1 + \alpha p_2)^2\right). \tag{55}$$

This completes the proof. □

A.3. Proof of Theorem 3

Theorem 3. *Let $\boldsymbol{q}, \boldsymbol{k} \in \{0,1\}^d$ be independent spike trains with firing rates $f_{\boldsymbol{q}}, f_{\boldsymbol{k}} \in (0,1)$ and $\alpha \in (0,1)$. Then the entropies satisfy: $H\left(\mathit{Sim}_{\alpha\text{-}\mathit{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right) > H\left(\mathit{Sim}_{\mathit{DP}}(\boldsymbol{q}, \boldsymbol{k})\right)$ and $H\left(\mathit{Sim}_{\alpha\text{-}\mathit{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right) > H\left(\mathit{Sim}_{\mathit{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right)$.*

Proof. From Theorem 2, the distribution of $\text{Sim}_{\text{Alpha}}(\boldsymbol{q}, \boldsymbol{k})$ is given by:

$$P\left(\text{score} = k + l\alpha\right) = \frac{d!}{k!\, l!\, (d-k-l)!} \cdot p_1^k\, p_2^l\, p_3^{d-k-l}, \tag{56}$$

for $0 \leq k \leq d$ and $0 \leq l \leq d - k$, where $p_1 = f_{\boldsymbol{q}} f_{\boldsymbol{k}}$, $p_2 = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})$, and $p_3 = 1 - p_1 - p_2$.

By Theorem 1, we know that $\text{Sim}_{\text{DP}}(\boldsymbol{q}, \boldsymbol{k})$ and $\text{Sim}_{\text{XNOR}}(\boldsymbol{q}, \boldsymbol{k})$ both follow binomial distributions:

$$\text{Sim}_{\text{DP}}(\boldsymbol{q}, \boldsymbol{k}) \sim \text{Bin}(d, p_1), \tag{57}$$

$$\text{Sim}_{\text{XNOR}}(\boldsymbol{q}, \boldsymbol{k}) \sim \text{Bin}(d, p_1 + p_2). \tag{58}$$

To compare entropies, we examine $H\left(\text{Sim}_{\text{Alpha}}(\boldsymbol{q}, \boldsymbol{k})\right)$ in relation to $H\left(\text{Sim}_{\text{DP}}(\boldsymbol{q}, \boldsymbol{k})\right)$ and $H\left(\text{Sim}_{\text{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right)$.

Observe that $\text{Sim}_{\text{Alpha}}(\boldsymbol{q}, \boldsymbol{k})$ is a function of two random variables, K and L, where K is the number of positions where $q_i = k_i = 1$ and L is the number of positions where $q_i = k_i = 0$. The joint distribution of (K, L) follows a multinomial distribution with parameters d and probabilities (p_1, p_2, p_3), where $p_3 = 1 - p_1 - p_2$. The entropy $H\left(\text{Sim}_{\text{Alpha}}(\boldsymbol{q}, \boldsymbol{k})\right)$ is equivalent to the joint entropy $H(K, L)$ because $\text{Sim}_{\text{Alpha}} = K + \alpha L$ is a deterministic function of K and L. Therefore,

$$H\left(\text{Sim}_{\text{Alpha}}(\boldsymbol{q}, \boldsymbol{k})\right) = H(K, L). \tag{59}$$

Next, consider the entropy $H\left(\text{Sim}_{\text{DP}}(\boldsymbol{q}, \boldsymbol{k})\right)$, which is the entropy of a binomial random variable $K \sim \text{Bin}(d, p_1)$:

$$H\left(\text{Sim}_{\text{DP}}(\boldsymbol{q}, \boldsymbol{k})\right) = H(K). \tag{60}$$

Similarly, the entropy $H\left(\text{Sim}_{\text{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right)$ corresponds to the entropy of $K + L$, which is a function of K and L:

$$H\left(\text{Sim}_{\text{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right) = H(K + L). \tag{61}$$

According to the chain rule for entropy, we have:

$$H(K, L) = H(K) + H(L \mid K). \tag{62}$$

As conditional entropy $H(L \mid K)$ is always non-negative ($H(L \mid K) \geq 0$), it follows that:

$$H(K, L) \geq H(K). \tag{63}$$

Equality holds if and only if L is a deterministic function of K, which is not the case here as $p_2 > 0$.

Furthermore, since $K + L$ is a deterministic function of K and L, the entropy satisfies:

$$H(K, L) \geq H(K + L). \tag{64}$$

Equality holds if and only if the function preserves all the uncertainty present in (K, L), which does not hold here due to the non-deterministic relationship between K and L.

Since both equalities cannot hold, combining these results we obtain:

$$H(K, L) > H(K) \quad \text{and} \quad H(K, L) > H(K + L). \tag{65}$$

Thus, the entropy of the Alpha similarity score exceeds that of both the Dot-Product and XNOR similarity scores:

$$\begin{aligned} &H\left(\mathrm{Sim}_{\mathrm{Alpha}}(\boldsymbol{q}, \boldsymbol{k})\right) > H\left(\mathrm{Sim}_{\mathrm{DP}}(\boldsymbol{q}, \boldsymbol{k})\right), \\ &H\left(\mathrm{Sim}_{\mathrm{Alpha}}(\boldsymbol{q}, \boldsymbol{k})\right) > H\left(\mathrm{Sim}_{\mathrm{XNOR}}(\boldsymbol{q}, \boldsymbol{k})\right). \end{aligned} \tag{66}$$

This completes the proof. □

A.4. The distribution of Attn' in α-SSA

The distribution of $\mathbf{Attn}'$ derived in α-SSA can be characterized by the following Theorem 4.

Theorem 4. *Let $\boldsymbol{v} \in \{0, 1\}^n$ be a spike train with firing rate $f_{\boldsymbol{v}}$, and let Score $\in \mathbb{R}^n$ be a vector of independent scores calculated using α-Similarity. For constants k and b, the attention output Attn' $= (k \cdot Score + b)^\top \boldsymbol{v}$ has an expected value*

$$\mathbb{E}[\textit{Attn'}] = n\left(kd\left(p_1 + \alpha p_2\right) + b\right) f_{\boldsymbol{v}} \tag{67}$$

and variance

$$\begin{aligned} \mathbb{V}[\textit{Attn'}] = n f_{\boldsymbol{v}}(1 - f_{\boldsymbol{v}})\Big(&k^2 d(p_1 + \alpha^2 p_2 \\ &- (p_1 + \alpha p_2)^2) + (kd(p_1 + \alpha p_2) + b)^2\Big). \end{aligned} \tag{68}$$

Proof. According to Corollary 1, each element of $\boldsymbol{Score}$ follows this distribution:

$$P(\mathrm{Score}_j = k + l\alpha) = \frac{d!}{k!\, l!\, (d - k - l)!} \cdot p_1^k\, p_2^l\, p_3^{d-k-l}, \tag{69}$$

where $0 \leq k \leq d$ and $0 \leq l \leq d - k$, with $p_1 = f_{\boldsymbol{q}} f_{\boldsymbol{k}}$, $p_2 = (1 - f_{\boldsymbol{q}})(1 - f_{\boldsymbol{k}})$, and $p_3 = 1 - p_1 - p_2$. The expectation and variance of each Score_j are given by:

$$\mathbb{E}[\mathrm{Score}_j] = d(p_1 + \alpha p_2), \tag{70}$$

$$\mathbb{V}[\mathrm{Score}_j] = d\left(p_1 + \alpha^2 p_2 - (p_1 + \alpha p_2)^2\right). \tag{71}$$

We now define the weighted sum:

$$\textit{Attn'} = (k \cdot \mathrm{Score} + b)^\top \boldsymbol{v} = \sum_{j=1}^{n} (k \cdot \mathrm{Score}_j + b) v_j. \tag{72}$$

To compute $\mathbb{E}[\textit{Attn'}]$, we use the linearity of expectation:

$$\begin{aligned} \mathbb{E}[\textit{Attn'}] &= \mathbb{E}\left[\sum_{j=1}^{n} (k \cdot \mathrm{Score}_j + b) v_j\right] \\ &= \sum_{j=1}^{n} \mathbb{E}[(k \cdot \mathrm{Score}_j + b) v_j]. \end{aligned} \tag{73}$$

Since each v_j is a Bernoulli random variable with success probability $f_{\boldsymbol{v}}$, we have

$$\mathbb{E}[(k \cdot \mathrm{Score}_j + b) v_j] = (k\mathbb{E}[\mathrm{Score}_j] + b)\, f_{\boldsymbol{v}}. \tag{74}$$

Substituting $\mathbb{E}[\mathrm{Score}_j] = d(p_1 + \alpha p_2)$, we obtain

$$\mathbb{E}[\textit{Attn'}] = n\left(kd(p_1 + \alpha p_2) + b\right) f_{\boldsymbol{v}}. \tag{75}$$

Next, we calculate $\mathbb{V}[\textit{Attn'}]$. Since *Attn'* is a sum of independent random variables, we can write

$$\mathbb{V}[\textit{Attn'}] = \sum_{j=1}^{n} \mathbb{V}((k \cdot \mathrm{Score}_j + b) v_j). \tag{76}$$

To compute $\mathbb{V}((k \cdot \mathrm{Score}_j + b) v_j)$, we apply the variance formula for the product of a random variable and a Bernoulli variable. Specifically, if Y is a Bernoulli random variable with success probability p, and X is an independent random variable, then:

$$\mathbb{V}(X \cdot Y) = \mathbb{E}[X^2] \cdot p \cdot (1 - p) + (\mathbb{E}[X])^2 \cdot p \cdot (1 - p). \tag{77}$$

Applying this formula, we get:

$$\begin{aligned} \mathbb{V}((k \cdot \mathrm{Score}_j + b) v_j) = \mathbb{E}[(k \cdot \mathrm{Score}_j + b)^2] f_{\boldsymbol{v}}(1 - f_{\boldsymbol{v}}) \\ + \left(\mathbb{E}[k \cdot \mathrm{Score}_j + b]\right)^2 f_{\boldsymbol{v}}(1 - f_{\boldsymbol{v}}). \end{aligned} \tag{78}$$

Then expand $\mathbb{E}[(k \cdot \mathrm{Score}_j + b)^2]$ and $\left(\mathbb{E}[k \cdot \mathrm{Score}_j + b]\right)^2$:

$$\mathbb{E}[(k \cdot \mathrm{Score}_j + b)^2] = k^2 \mathbb{V}[\mathrm{Score}_j] + (k\mathbb{E}[\mathrm{Score}_j] + b)^2, \tag{79}$$

$$\left(\mathbb{E}[k \cdot \mathrm{Score}_j + b]\right)^2 = (kd(p_1 + \alpha p_2) + b)^2. \tag{80}$$

Substituting these into the expression for $\mathbb{V}((k \cdot \mathrm{Score}_j + b) v_j)$, we obtain:

$$\begin{aligned} \mathbb{V}((k \cdot \mathrm{Score}_j + b) v_j) = f_{\boldsymbol{v}}(1 - f_{\boldsymbol{v}})\Big(&k^2 d(p_1 + \alpha^2 p_2) \\ &- k^2 d(p_1 + \alpha p_2)^2 \\ &+ (kd(p_1 + \alpha p_2) + b)^2\Big). \end{aligned} \tag{81}$$

Summing over all j, we get the total variance of *Attn'*:

$$\begin{aligned} \mathbb{V}[\textit{Attn'}] = n f_{\boldsymbol{v}}(1 - f_{\boldsymbol{v}})\Big(&k^2 d(p_1 + \alpha^2 p_2 \\ &- (p_1 + \alpha p_2)^2) + (kd(p_1 + \alpha p_2) + b)^2\Big). \end{aligned} \tag{82}$$

This completes the proof. □

Table 4. Correlation of α and metrics across different α-SSA layers

α	α-SSA Layer	$f_{\mathbf{Q}}$ (%)	$f_{\mathbf{K}}$ (%)	$f_{\mathbf{V}}$ (%)	$f_{\mathbf{Attn}}$ (%)	k	b	Acc. (%)
0	1	31.5	9.7	10.2	35.1	0.45	-0.36	79.04
	2	33.2	8.5	12.8	33.2	0.27	-0.44	
	3	29.8	9.9	13.4	32.1	0.34	-0.35	
	4	27.4	10.8	12.0	28.2	0.55	-0.22	
0.2	1	31.9	16.6	11.2	19.8	0.18	-0.90	79.81
	2	32.5	19.9	13.8	23.6	0.16	-0.84	
	3	37.7	20.6	12.5	20.5	0.17	-0.90	
	4	34.6	22.0	12.0	14.1	0.17	-0.92	
0.4	1	32.8	28.9	10.6	26.3	0.15	-1.37	79.88
	2	23.4	24.7	13.6	22.5	0.13	-1.22	
	3	23.2	23.0	11.9	18.9	0.14	-1.31	
	4	35.0	30.1	10.0	14.1	0.15	-1.42	
0.6	1	22.4	25.1	11.0	24.2	0.14	-1.76	79.63
	2	14.9	22.6	12.4	21.3	0.13	-1.74	
	3	12.0	13.7	13.0	18.5	0.12	-1.85	
	4	11.6	16.9	9.7	12.4	0.14	-2.09	
0.8	1	14.6	19.5	10.2	29.7	0.14	-2.59	79.42
	2	13.5	23.4	13.0	32.3	0.12	-2.22	
	3	14.1	26.7	10.6	28.9	0.14	-2.36	
	4	9.2	22.0	8.4	17.9	0.16	-2.91	
1	1	14.5	20.8	11.0	33.4	0.15	-3.31	79.32
	2	12.3	20.0	13.6	32.6	0.13	-3.12	
	3	9.5	13.6	13.4	31.5	0.15	-3.80	
	4	4.2	12.9	9.4	23.7	0.20	-5.27	

Table 5. Main configurations of models in image classification experiments.

Type	Stage	Tokens	Layer Specification	Dimensions
ViT	1	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	384 / 512
	2		Spiking Transformer Encoder Block	
Swin	1	$\frac{H}{4} \times \frac{W}{4}$	Downsampling Conv-based SNN Block	96 / 128
	2	$\frac{H}{8} \times \frac{W}{8}$	Downsampling Conv-based SNN Block	192 / 256
	3	$\frac{H}{16} \times \frac{W}{16}$	Downsampling Spiking Transformer Encoder Block	384 / 512

B. Spiking Firing Rates in SNNs

The spiking firing rate in SNNs is related to their training methods. Based on the training approach, SNNs can be broadly classified into two categories: ANN-to-SNN conversion and direct training. SNNs obtained through ANN-to-SNN conversion require mapping the activation values of the ANN to firing rates, which typically results in relatively high firing rates, around 50%. In contrast, directly trained SNNs can naturally optimize the firing rate to be significantly lower, further reducing power consumption.

In recent research, spiking Transformer models predominantly employ direct training methods, demonstrating significant sparsity. Table 6 presents the evaluated firing rates of spiking neurons across four layers of the SSA module in the Spikformer model on the CIFAR-100 dataset. Here, Q,

Table 6. Firing Rates in SSA of Spikformer on CIFAR100

SSA Layer	Q	K	V	Attn	Output
1	22.9	9.3	11.3	28.4	8.7
2	19.1	8.6	8.7	23.3	9.8
3	19.7	7.0	8.9	21.0	11.6
4	21.9	6.0	7.8	5.2	20.2

K, V, Attn, and Output correspond to the spiking neurons generating $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, $\mathbf{Attn}$, and $\mathbf{Z}$ described in Sec. 3.3, respectively. The table shows that, while firing rates vary across different neurons and layers, they consistently remain below 30%. This indicates that the spiking data is highly sparse, with a substantial proportion of non-spikes.

According to information theory, rare events carry more information, meaning that each spike, which occurs with low probability, conveys a significant amount of information. Therefore, it is reasonable to consider spikes in spiking data as more important than non-spikes, prompting us to differentiate the distinct significance of spikes from non-spikes in spiking self-attention.

C. Layer-wise Details for Study of α

In Sec. 5.2, we analyze the influence of α on the key metrics and parameters within the α-SSA module. The experiments are conducted on the Spikformer architecture on CIFAR100 and the metrics include the linear transformation factors k and b; the spiking firing rates $f_{\mathbf{Q}}$, $f_{\mathbf{K}}$, $f_{\mathbf{V}}$, and $f_{\mathbf{Attn}}$ responsible for generating $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and the spiking self-attention output $\mathbf{Attn}$, respectively; and the network's performance accuracy (Acc.). Here, we present the detailed data across different α-SSA layers, shown in Table 4. As α increases from 0 to 1, both the spiking firing rates and the parameters of linear transformations are influenced and undergo changes, with slight variations observed across different layers. The table shows that as the network depth increases, the magnitude of changes in the respective metrics becomes greater, suggesting that α has a more significant impact on the deeper layers.

D. Architecture Details in Experiments

In our image classification experiments, we evaluate the proposed α-SSA module on spiking Transformers using both ViT and Swin Transformer architectures. The details are summarized in Table 5. For the ViT architecture, the model consists of a patch embedding layer followed by several spiking Transformer encoder blocks. And the model with Swin Transformer architecture comprises three stages, each beginning with a downsampling layer and then followed by either a convolution-based SNN Block or multiple spiking Transformer encoder blocks.